(19) Europäisches Patentamt

European Patent Office

Office européen d s brevets

(11) **EP 0 786 132 B1**

(12) **EUROPEAN PATENT SPECIFICATION**

(45) Date of publication and mention
of the grant of the patent:
26.04.2000 Bulletin 2000/17

(21) Application number: **96924104.1**

(22) Date of filing: **02.08.1996**

(51) Int Cl.7: **G10L 13/04**

(86) International application number:
PCT/IB96/00770

(87) International publication number:
WO 97/07499 (27.02.1997 Gazette 1997/10)

(54) **A METHOD AND DEVICE FOR PREPARING AND USING DIPHONES FOR MULTILINGUAL TEXT-TO-SPEECH GENERATING**

VERFAHREN UND VORRICHTUNG ZUR BEREITSTELLUNG UND VERWENDUNG VON DIPHONEN FÜR MEHRSPRACHIGE TEXT-NACH-SPRACHE SYSTEME

PROCEDE RELATIF A L'ELABORATION ET A L'UTILISATION DE DIPHONEMES POUR LA SYNTHESE MULTILINGUE DE TEXTE EN PAROLE ET DISPOSITIF CORRESPONDANT

(84) Designated Contracting States:
DE ES FR GB IT

(30) Priority: **14.08.1995 EP 95202202**

(43) Date of publication of application:
30.07.1997 Bulletin 1997/31

(73) Proprietor: **Koninklijke Philips Electronics N.V.**
5621 BA Eindhoven (NL)

(72) Inventor: **COLLIER, René, Pierre, Gaston**
NL-5621 BA Eindhoven (NL)

(74) Representative: **Peters, Rudolf Johannes**
INTERNATIONAAL OCTROOIBUREAU B.V.,
Prof. Holstlaan 6
5656 AA Eindhoven (NL)

(56) References cited:
EP-A- 0 262 938   **EP-A- 0 694 904**
WO-A-96/12271

- SPEECH TECH. '85, Volume 1, No. 2, 1985, (New York), G.L. CAMPOS, "Speech Synthesis for Unrestricted Vocabulary: The Phoneme-Diphone Approach", pages 307-309.
- PROCEEDINGS OF SPEECH '88, 1988, (Edinburgh), M. ONADOU et al., "A Mutilanguage Speech Synthesis System", pages 57-58.

EP 0 786 132 B1

1   EP 0 786 132 B1   2

Description

FIELD OF THE INVENTION

[0001] The invention relates to a method for preparing diphones for text-to-speech generating by selectively pronouncing a set of selected diphones, processing each such diphone for persistent storage, and storing each processed diphone in an individually addressable storage location. Diphones have been found an excellent vehicle for speech decomposition and representation. For example, Patents US 4,813,074 to Marcus and US 5,109,418 to Van Hemert, both assigned to the present assignee teach the extraction of segment boundaries from spoken text. These methods can be excellently used for constituting a diphone from a pair of successive segmented elements of predetermined lenghts before and after the boundary, respectively. A set of such diphones in turn allows for representing speech in a robust manner. In particular, subsequent speech rendering is elementary, because no complex rules are necessary for realising the coarticulation between contiguous phonemes. Once generated, the diphones are stored in conventional storage for later usage. The English language may be well represented by using a 1500 diphone repertory. The document Proceedings of Speech '88, pages 55-59, discloses a multilanguage speech synthesis system which is employed in a text-to-speech system. The hardware realisation employs two diphone dictionaries which are stored in two ROM's, respectively. The present inventor has encountered the need for representing multilingual speech and in particular has found the required storage at 1500 locations for each language extremely burdening, especially in an environment where minimal cost is a prerequisite. Such environments up to now usually have done either without speech rendering, or with only a single language, or with low-quality rendering methods, where storage cost is often irrelevant.

SUMMARY TO THE INVENTION

[0002] Accordingly, amongst other things it is an object of the present invention to provide a method according to the preamble that allows to generate a diphone repertory for multilingual speech rendering that alleviates storage capacity requirements. Now, according to one of its aspects, the invention is characterized in that amongst such set as spoken by a single person on a homophony basis assigning each diphone to one or more diverse languages for diminishing storage requirements through sharing of selective diphones amongst more than one language. For example, six of the main European languages can be collectively represented by a set of only some 3000 diphones. This is a reduction by a factor of three versus the sixfold enlargement that a sixfold repetition of the 1500 diphones of each particular single language would require. The invention relies on the combined idea, first, that many languages, in particular European languages, have identical or nearly identical combinations of successive sounds, and second, that certain persons have the ability to pronounce such combinations in a way that would be appropriate to several particular languages, even if the person in question were not able to speak the language. Often, such persons are multilingual. In fact, it were only necessary to correctly pronounce a single combination of two sounds to the satisfaction of another person, who is able to speak the 'new' language. The more or less identical pronunciation of a particular diphone in various languages need not be based on correspondence in orthography. On the other hand, identical orthography need not go hand in hand with identical pronunciation.

[0003] The above implies that the reusable diphones are always produced by the same voice. On the other hand, it would be fully unacceptable if speech is constructed from a sequence of diphones that are diverse in origin in that they have been generated by multiple persons. The storage proper can be effected in a software-controlled manner: it can be in the form of speech segments that are represented by an appropriate waveform, or rather by parametrized data. The field of application of the invention can be various, such as traffic information systems, public transport enquiry systems, and various others.

[0004] Advantageously, certain diphones are stored with processing qualifiers that discriminate processing operations on the diphone in question amongst such various languages sharing that diphone. Such extra processing qualifiers allow to further diminish storage requirements. In this way, also diphones that have particular limited differences can be shared amongst various languages. Such differences may pertain to duration of vowels, aspiration, or vowel quality.

[0005] The invention also relates to a method for text-to-speech transforming and to a device therefor. One particular advantage of the invention is the easy manner that exonyms can be inserted into a spoken text. This may, for example, be realized by inserting language identifiers at the start and the termination of the exonym. Further advantageous aspects of the invention are recited in dependent Claims.

BRIEF DESCRIPTION OF THE DRAWING

[0006] These and other aspects and advantages of the invention will be explained in detail hereinafter with reference to the disclosure of preferred embodiments, and in particular with reference to the appended Figures that show:

Figure 1, a flow chart of the diphone preparing;
Figure 2, an apparatus for text-to-speech transformation.

3 · EP 0 786 132 B1 4

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

[0007] Figure 1 shows a flow chart of the diphone preparing. In block 20, the system and user persons prepare for the task. Appropriate storage and processing facilities are allocated, and a list of diphones is set up. In block 22, the first diphone is presented to the single person that is going to pronounce all diphones of the ultimate set. The presentation may be in visual or audio form. In block 24, the first diphone is pronounced by the single person. In block 26, the sound is checked for correct pronunciation. This can be done in a subjective manner by one or more reviewer persons, or alternatively, by some kind of machine comparison with renderings of the same diphone by a standard voice. The pronunciation may be done in an isolated manner, as a sequence of intendedly identical diphones, or in another appropriate way. In case of plural utterances, the 'best' one may be selected by the reviewer. If the pronunciation is considered insufficient, the system goes back to block 24 and requires another pronunciation of the same diphone. The dialog may be supported by visual or audio stimuli, in a manner known for human-machine dialogs in general. For simplicity this is not detailed further.

[0008] If the pronunciation is considered sufficient, in block 28 the sound is processed, such as clipped in time domain, standardized in volume, filtered, if necessary transformed into a set of parameters, and stored in a conventional manner. For ultimate usage, this storage must be random-access, but at this particular instant such is not yet necessary. A diphone identifier must be added in that case. In block 30 the system detects whether it is ready. If no, it reverts to block 22 for the next diphone. If yes, in block 32 the activities are closed, such as by outputting the results, to wit now a compact set of diphones that is suitable for outputting of speech in more than one language. Likewise, the output may relate to an enhancing set of diphones for expanding an n-language set to an n+1-language set.

[0009] For certain applications, the stored diphones may have processing qualifiers that discriminate processing operations on the diphone in question amongst such various languages sharing this diphone. Examples thereof will be discussed hereinafter.

[0010] Figure 2 is a block diagram of an apparatus for text-to-speech transformation. Block 48 represents the storage of the multilingual set, for example realized in semiconductor technology. In the embodiment, the device is used for four different languages, as symbolized by blocks 40-46. Diphone block 40 relates to language A, block 42 relates to language B, block 44 relates to language C, and block 46 relates to language D. In the principal storage, the topmost block therefore relates to languages A and B, the second block to languages A, C, and D, and so on. In the example, only the second last diphone relates only to language C, one diphone relates to all four languages, and the others to various

different combinations. The distribution depends on the set of languages, and to a certain extent is open to choice. Generally, the number of different diphones is a matter of choice. In various cases, a particular diphone may be reused among several languages, but the usage of a unique diphone for each language would give a slight improvement. The overall result is a matter of trade-off.

[0011] Block 56 represents the input of a string of diphone codes. These may be associated with a duration, aspiration, prosody, or other processing qualifier indication that need not unique for the same diphone, either when in different languages, or in various positions in a sentence. By itself, differentiation of prosody versus the position in a sentence has been disclosed in prior, non-prepublished European Patent Application EP-A-0 774 152. The above disclosure however, stores similar speech items for one language in different ways. The present invention goes the other way round by sharing storage among various languages. Likewise, prosody control signals may accompany the diphone string. Line 52 transfers the address code to the storage organization 48. The latter may be location-based or content-addressed, such as alphabetical. Bracket 50 symbolizes the read-out mechanism of the storage proper. Block 58 is the outputting mechanism of the storage, such as amplifier or hold that is controlled by line 54, such as by a synchronizing signal. Block 60 effects the transformation of the diphone for subsequent rendering, such as by windowing in a sequence of diphones, or by translating a parametrized diphone to the 'real' sound. Block 62 is the audio output mechanism, such as a combination of output amplifier and loudspeaker. Item 64 symbolizes the audio output that may be personal or public address. If necessary, speaker verification can be used to verify that all diphones so outputted indeed stem from a single speaker person.

FURTHER CONSIDERATIONS

[0012] A reference for rendering speech in a multi-language environment can be found in R. Collier, Multi-language Intonation Synthesis, Journal of Phonetics 19 (1991), 61-73.

[0013] Further, the processing of the diphones has two aspects. In the first place, there is the processing on the level of the utterance, that is, the diphones are concatenated and subsequently, speech pitch and rhythm are added. In the second place, there is the processing on the level of the diphone itself, before it is entered into an utterance. The latter kind of processing should lead to further reduction of storage requirements. The following are a few examples:

1. The English word 'mean', the Dutch word 'Mien' (girl's name), and the French word 'mine' (mine, such as for producing coal) share the vowel, the only difference for near-perfect pronunciation being its

5   EP 0 786 132 B1   6

duration, which is longest in English, shortest in French. The policy therefore is to use the same combination of stored diphones (mi + in), but the storage of the first diphone carries different duration indications for each respective language that may be in the range of 200-80 milliseconds.

2. Even in the same language, a particular diphone may have to be rendered in different ways. For example, the english word 'pin' has aspiration on the sound 'p', which the word 'spin' has not. The same diphone (pi) can be used in other languages, either with, or without aspiration. A particular manner for diminishing storage requirements is to store the diphone as a waveform with aspiration, and to insert marker indications at the start and at the end of the aspiration sound. If required, the aspiration can then be foregone. In both of the above cases, the processing qualifiers apply differentially to the time domain.

3. If the diphones are rather stored as parametric information, the processing can be made differentially in the spectral domain. In the Dutch word 'idee' (idea), the second vocal has more or less the character of a diphthong. In French, the word 'idle' does not have this property. By itself, the lending of the 'diphthong' character can be effected by formant manipulation, known by itself. The Dutch representation of the diphone in question can now be associated with a processing qualifier that controls this manipulation, thereby necessitating the storage of only one diphone. In a similar manner, the diphthong aspect can have different gradations in intensity, such as in the case of 'dose' (English), 'dozen' (Dutch: boxes), and 'dose' (French: dose). It is well known that the number of relevant formants is small, and their frequencies can be stored as a single quantity, or as a slope between two different quantities. In the LPC-based approach, the spectral manipulation can be combined with manipulation in the time domain.

4. However, even without these differentiating processing operations, the approach of the invention remains valid, either with a slightly lower sound quality, or with a somewhat larger storage. The following examples are given, wherein standard phonetic conventions have been used. For simplicity, words have been used that have (nearly) the same text representation in the various languages. However, the same reusability principle applies when the words are phonetically corresponding only in part.

diphone number	representation	language
1	#s	NL, ENG
2	sp	NL, ENG
3	pl	NL, ENG
4	ln	NL, ENG
5	n#	NL, ENG

[0014]   In the above, # indicates silence at the beginning and end, respectively, of the word in question. In both languages, the combination of diphones 1+2+3+4+5 renders the word **'spin'**.

6	#p^h	ENG, DE
7	p^{h}i	ENG, DE
8	ik(#)	ENG, DE

[0015]   The sequence 6+7+8 renders the word 'p^{h}ik' (peak, Pik) in English and German, respectively.

9	#i	NL, DE, FR
10	id	NL, DE, FR
11	de(#)	NL, DE, FR
12	de^i(#)	NL

[0016]   The sequence 9+10+11 renders the word 'ide' (German 'Idee', French 'idée'), the sequence 9+10+12 renders the word 'ide^1' (Dutch 'idee', with a slight diphthong aspect).

13	#f	NL, DE, FR
14	fa	NL, DE, FR
15	at	NL, DE, FR
16	ta	NL, DE, FR
17	al	NL, FR, DE*
18	l#	NL, FR, DE*

[0017]   The sequence 13+14+15+16+17+18 renders the basis of the Dutch word 'fataal' (fatal), German 'fatal' or French 'fatal'. For each language the both vocals get their respective lengths adjusted. If required, the last two diphones are stored another time for German only (signalled by an asterisk *) with a somewhat 'thinner' pronunciation of the *l*, because it is difficult to achieve this with differentiated processing for this language.

Claims

1.   A method for preparing diphones for text-to-speech generating by selectively pronouncing a set of selected diphones, processing each such diphone for persistent storage, and storing each processed diphone in an individually addressable storage loca-

7   EP 0 786 132 B1   8

tion, characterized by amongst such set as spoken by a single person on a homophony basis assigning each diphone to one or more diverse languages for diminishing storage requirements through sharing of selective diphones amongst more than one language.

2. A method as claimed in Claim 1, wherein certain diphones are stored with processing qualifiers that discriminate processing operations on the diphone in question amongst such various languages sharing that diphone.

3. A method for multi-language text-to-speech transforming by receiving said text as a string of diphone identifiers, read-accessing by each identifier a diphone storage location, and converting each diphone information so read out in a code-to-sound converter to an associated sound item, characterized in that such diphones have been generated according to the method of Claims 1 or 2, and selective ones amongst said locations are each responsive to an associated plurality of diphone identifiers that are assigned to a plurality of languages.

4. A method as claimed in Claim 3, furthermore accessing certain diphones that are stored with processing qualifiers for discriminating processing operations on the diphone in question amongst such various languages sharing that diphone, and wherein for speech rendering in a particular language only qualifiers for that particular language are taken into account.

5. A device for multi-language text-to-speech transforming having input means for receiving said text as a string of diphone identifiers, read-access means for accessing by each identifier an associated diphone storage location, and converting means fed by said storage means for converting each diphone information so read out in a code-to-sound converter to an associated sound item, characterized in that such diphones have been generated according to the method of Claims 1 or 2 and selective ones amongst said locations are each responsive to an associated plurality of diphone identifiers that are shared amongst a plurality of languages.

6. A device as claimed in Claim 5, furthermore having processing means for accessing diphones that are stored with processing qualifiers for discriminating processing operations on the diphone in question amongst such various languages sharing that diphone, and wherein for speech rendering in a particular language such processing means is only responsive to qualifiers for that particular language.

Patentansprüche

1. Verfahren zum Vorbereiten von Diphonen zur Text-in-Sprache-Generierung durch selektives Aussprechen einer Menge selektierter Diphone, Verarbeiten jedes solchen Diphons für dauernde Speicherung und Speichern jedes verarbeiteten Diphons in einer einzeln adressierbaren Speicherstelle, gekennzeichnet durch Zuweisen jedes Diphons innerhalb einer solchen von einer einzelnen Person gesprochenen Menge auf Homophoniebasis an eine oder mehrere verschiedene Sprachen zum Verringern der Speicheranforderungen durch Teilen selektiver Diphone unter mehr als einer Sprache.

2. Verfahren nach Anspruch 1, wobei bestimmte Diphone mit Verarbeitungskennzeichnern gespeichert werden, die Verarbeitungsoperationen an dem betreffenden Diphon innerhalb solcher verschiedenen dieses Diphon teilenden Sprachen unterscheiden.

3. Verfahren für mehrsprachiges Text-in-Sprache-Transformieren durch Empfangen des genannten Textes als Kette von Diphonkennungen, Lesezugriff jeder Kennung auf eine Diphon-Speicherstelle und Umsetzen jeder so ausgelesenen Diphoninformation in einem Code-Laut-Umwandler in ein zugehöriges Laut-Item, dadurch gekennzeichnet, dass derartige Diphone gemäß dem Verfahren der Ansprüche 1 oder 2 generiert worden sind und selektive Stellen innerhalb dieser Stellen jeweils auf eine zugehörige Vielzahl von Diphonkennungen ansprechen, die einer Vielzahl von Sprachen zugeordnet sind.

4. Verfahren nach Anspruch 3, weiterhin Zugreifen auf gewisse Diphone, die mit Verarbeitungskennzeichnern gespeichert sind zum Unterscheiden von Verarbeitungsoperationen an dem betreffenden Diphon innerhalb solcher verschiedenen dieses Diphon teilenden Sprachen, und wobei zur Sprachwiedergabe in einer speziellen Sprache nur Kennzeichner für diese spezielle Sprache berücksichtigt werden.

5. Einrichtung für mehrsprachiges Text-in-Sprache-Transformieren mit Eingabemitteln zum Empfangen von Text als Kette aus Diphonkennungen, Lesezugriffmitteln zum Zugreifen jeder Kennung auf eine zugehörige Diphonspeicherstelle und von den genannten Speichermitteln gespeisten Umwandlungsmitteln zum Umwandeln jeder so ausgelesenen Diphoninformation in einem Code-Laut-Wandler in ein zugehöriges Laut-Item, dadurch gekennzeichnet, dass derartige Diphone gemäß dem Verfahren der Ansprüche 1 oder 2 generiert worden sind und selektive Stellen innerhalb dieser Stellen

9   EP 0 786 132 B1   10

jeweils auf eine zugehörige Vielzahl von Diphon-
kennungen ansprechen, die unter iner Vielzahl
von Sprachen geteilt werden.

6. Einrichtung nach Anspruch 5, weiterhin mit Verar-
beitungsmitteln zum Zugreifen auf Diphone, die mit
Verarbeitungskennzeichnern gespeichert sind zum
Unterscheiden von Verarbeitungsoperationen an
dem betreffenden Diphon innerhalb solcher ver-
schiedenen dieses Diphon teilenden Sprachen, und
wobei zur Sprachwiedergabe in einer speziellen
Sprache solch Verarbeitungsmittel nur auf Kenn-
zeichner für diese spezielle Sprache anspricht.

Revendications

1. Procédé relatif à l'élaboration de diphonèmes pour
la synthèse de texte en parole en prononçant de
manière sélective une série de diphonèmes sélec-
tionnés, en traitant chacun de ces diphonèmes en
vue de leur mémorisation persistente, et en mémo-
risant chaque diphonème traité à un emplacement
de mémoire apte à être adressé de manière indivi-
duelle, caractérisé par l'attribution, parmi cette série
telle que prononcée par une seule personne, sur la
base d'une homophonie, de chaque diphonème à
une ou plusieurs langues pour diminuer les exigen-
ces de mémoire en partageant des diphonèmes sé-
lectionnés entre plus d'une langue.

2. Procédé selon la revendication 1, dans lequel cer-
tains diphonèmes sont mémorisés avec des quali-
ficatifs de traitement qui établissent une distinction
entre des opérations de traitement sur le diphonè-
me en question parmi les diverses langues parta-
geant ce diphonème.

3. Procédé de transformation multilingue de texte en
parole par réception dudit texte sous forme d'une
chaîne d'identificateurs de diphonèmes, par acces-
sion en lecture, par chaque identificateur, à un em-
placement de mémorisation de diphonème, et par
conversion de chaque information de diphonème
ainsi lue dans un convertisseur code/son en un élé-
ment sonore associé, caractérisé en ce que ces di-
phonèmes ont été produits suivant le procédé de la
revendication 1 ou 2 et que des emplacements sé-
lectionnés parmi lesdits emplacements réagissent
chacun à une pluralité associée d'identificateurs de
diphonèmes qui sont attribués à une pluralité de
langues.

4. Procédé selon la revendication 3, accédant en
outre à certains diphonèmes qui sont mémorisés
avec les qualificatifs de traitement pour distinguer
les opérations de traitement sur le diphonème en
question parmi les diverses langues partageant ce

diphonème, t dans lequel, pour la reproduction
des paroles dans une langue en particulier, seuls
les qualificatifs pour cette langue en particulier sont
pris en compte.

5. Dispositif pour la transformation multilingue de tex-
te en parole ayant un moyen d'entrée pour recevoir
ledit texte sous la forme d'une chaîne d'identifica-
teurs de diphonèmes, un moyen d'accès en lecture
pour que chaque identificateur accède à un empla-
cement associé de mémorisation de diphonème et
un moyen de conversion alimenté par ledit moyen
de mémoire pour convertir chaque information de
diphonème ainsi lue dans un convertisseur code/
son en un élément sonore associé, caractérisé en
ce que ces diphonèmes ont été produits suivant le
procédé de la revendication 1 ou 2 et en ce que les
emplacements sélectionnés parmi lesdits emplace-
ments réagissent chacun à une pluralité associée
d'identificateurs de diphonèmes qui sont partagés
entre une pluralité de langues.

6. Dispositif selon la revendication 5, ayant en outre
un moyen de traitement pour accéder aux diphonè-
mes qui sont mémorisés avec les qualificatifs de
traitement destinés à opérer une distinction entre
les opérations de traitement sur le diphonème en
question parmi les diverses langues partageant ce
diphonème, et dans lequel, pour la reproduction de
la parole dans une langue en particulier, ce moyen
de traitement réagit uniquement aux qualificatifs de
cette langue en particulier.

EP 0 786 132 B1

```
        ( START )─── 20
             │
             ▼
      ┌──────────────┐
      │   PRESENT    │─── 22
      └──────────────┘
             │
             ▼
      ┌──────────────┐
      │  PRONOUNCE   │─── 24
      └──────────────┘
             │
             ▼
          ╱────╲─── 26
         ╱ O.K.? ╲
         ╲       ╱
          ╲────╱
             │
             ▼
      ┌──────────────┐
      │ PROCESS/STORE│─── 28
      └──────────────┘
             │
             ▼
          ╱────╲─── 30
         ╱ LAST ? ╲
         ╲       ╱
          ╲────╱
             │
             ▼
        ( STOP )─── 32
```
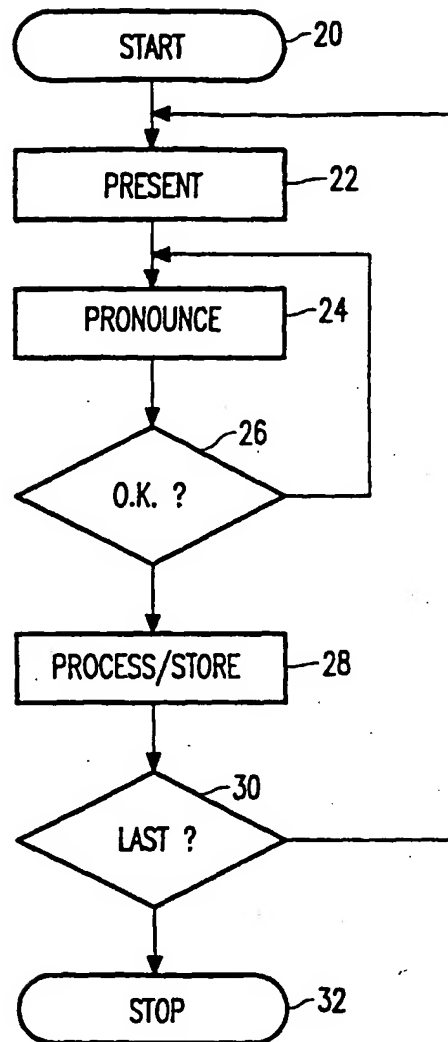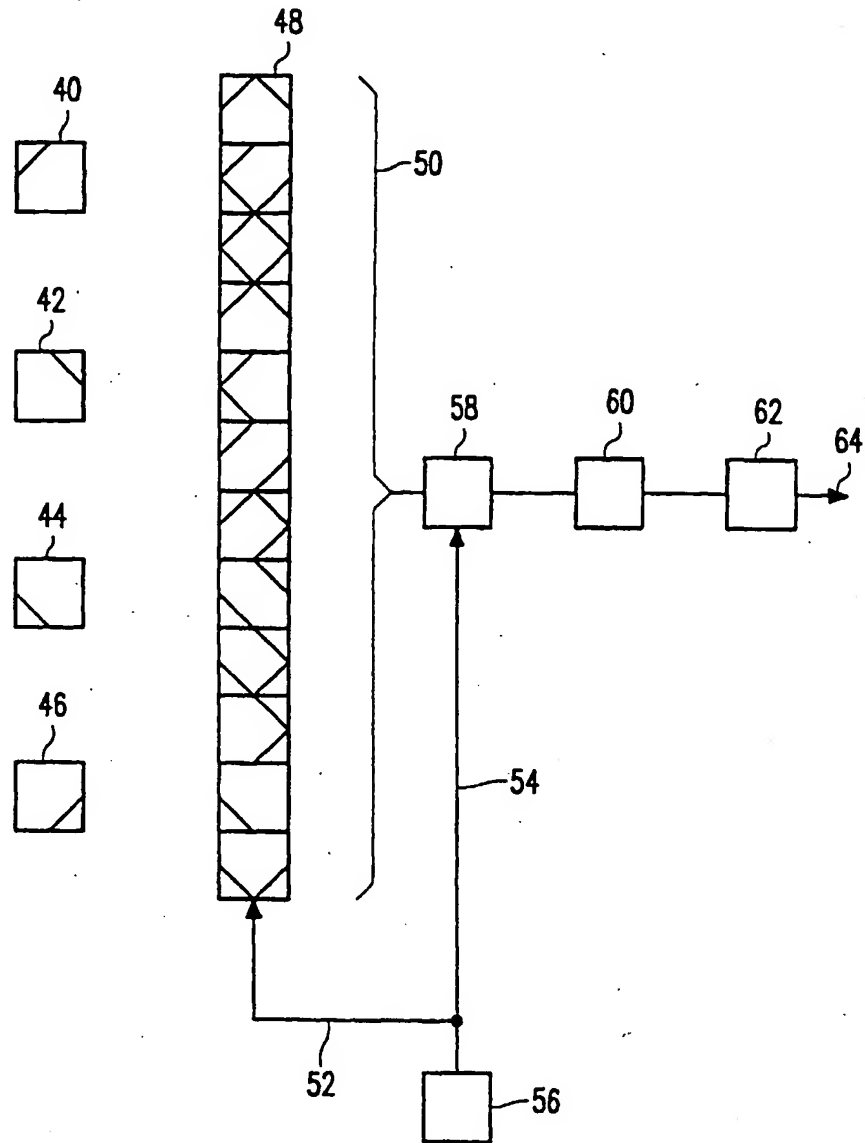
FIG. 1

EP 0 786 132 B1

FIG. 2